Video Object Segmentation Introducing Depth and Motion Information

Montse Pardàs
Universitat Politècnica de Catalunya, UPC
Barcelona, SPAIN
montse@gps.tsc.upc.es

ABSTRACT

In the paper we present a method to estimate relative depth between objects in scenes of video sequences. The information for the estimation of the relative depth is obtained from the overlapping produced between objects when there is relative motion as well as from motion coherence between neighbouring regions. A relaxation labelling algorithm is used to solve conflicts and assign every region to a depth level. The depth estimation is used in a segmentation scheme which uses grey level information to produce a first segmentation. Regions of this partition are merged on the basis of their depth level.

1. Introduction

Video sequence description in terms of objets is becoming an increasing area of interest. New standards like MPEG-4 and MPEG-7 will require this kind of description for different applications like video coding, functionalities as video composing and for indexing video data bases. In this context, several segmentation schemes are being studied. They are based on different types of information, mainly grey level [3] and motion [1] homogeneity.

In [5] we proposed to use an additional information for the segmentation, the relative depth of the objects of the scene. In that work, a bottom-up segmentation was proposed. First, a partition composed of regions homogeneous in grey level was produced. Then, these regions were merged on the basis of the depth level to which they belonged. The monocular depth estimation was based on the motion of the regions which produced occlusions between them. These occlusions were found analysing the evolution of the grey level regions.

In this paper we propose to use motion coherence information together with the depth level information in the top level of the segmentation scheme. This allows to use smaller regions in the bottom level of the segmentation, thus obtaining finer contours (the bottom level is based on grey level homogeneity), and also to solve uncertainties which were produced when there was not occlusion information between objects. The occlusion information defines the relative depth between pairs of neighbouring regions. In this case, besides the evolution of the grey level regions used in [5], the texture information is introduced in the study of the occlusions, in order to have more reliable information. Afterwards, a relaxation algorithm is used in order to assign a level of depth to every region.

2. General segmentation structure

In order to be suitable for coding or object manipulation applications, a segmentation must be continuous along the time dimension [7]. That is, not only we must avoid abrupt changes in the partition for successive images, but we also have to define the correspondence between regions of the different images. For the creation of such a segmentation we propose a two-level bottom-up approach. The bottom level is based only on the grey level information [3], and it is described in section 3, while the top level uses the relative depth of the regions in order to merge regions from the previous level. This step is described in section 4. Temporal continuity is achieved by means of a region tracking procedure implicit in the grey level segmentation [4] and by filtering the depth estimation.

3. Bottom level segmentation

In [4] we proposed a time-recursive segmentation scheme relying on the grey level information. This scheme produces a continuous segmentation along the time dimension, by producing a first segmentation of the initial image and following the regions with a region tracking procedure implicit in the segmentation. In the present work, the bottom level of the two-level

0-8186-8821-1/98 $10.00

segmentation is obtained using this scheme. In this segmentation, regions are defined by their grey level homogeneity. Two modes of operation can be distinguished: intra-frame and inter-frame. The whole process is illustrated in Figure 1.

3.1 Intra-frame

The objective of this mode is to produce an initial segmentation for the first image of the sequence (or whenever a refreshment is required). It is a top-down procedure which first produces a coarse segmentation with a reduced number of regions. Then, it is progressively improved by introducing more regions. Morphological tools are used to create this hierarchy of partitions. The properties of these tools allow to produce in the coarser levels regions with perfectly located contours. So, in the finer levels these contours do not have to be modified, and only new regions are introduced. Each segmentation level involves four basic steps:

- Image modelling. The partition obtained in the corresponding upper level (the whole image in the first level) is modelled and substracted from the original image in order to create the coding residue. This residue contains information of the regions which cannot be correctly represented and should be further segmented.
- Image simplification. The coding residue is simplified in order to eliminate the information which will not be used for segmentation in the current level. Morphological filters are used for this aim. Two different simplification criteria are generally used: size and contrast of the regions.
- Marker extraction. The goal of this step is to produce markers identifying the interior of the regions that will be segmented.
- Decision. The precise contours of the regions identified are located. The morphological watershed is used.

3.2 Inter-frame

The intra-frame procedure produces, at the lower level, a partition of the image where regions are homogeneous in grey level. This partition defines the segmentation of the bottom level of our general structure. The inter-frame procedure follows these regions along the sequence and introduces new regions when necessary. A time-recursive method is used. That is, the partition of every image is computed taking as starting point the partition of the previous image. For this aim, the previous partition is projected, and then the new regions corresponding to the current frame are extracted.

- Partition projection. First, motion between the previous frame and the current one is estimated. Then, the previous partition is motion compensated, and a watershed procedure is used to find the right contours of the previous regions in the current frame.
- Extraction of new regions. The same procedure explained in the intra-frame mode is used to extract new regions corresponding to appearing objects.

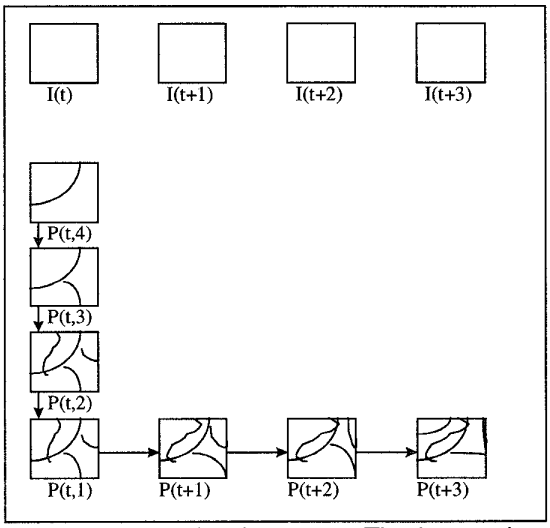

Figure 1. Bottom level process. The bottom level partition of the images I(t) to I(t+3) corresponds to partitions P(t,1) to P(t+3). The first one (P(t,1)) is obtained with the intra-frame mode, which consists in a grey level hierarchical segmentation process. The following ones are obtained with the inter-frame mode, which consists in a time recursion of projection and extraction of new regions.

4. Top level: Relative depth estimation

In this level a classification of the regions obtained in the previous level is done relying on an estimation of the relative depth between these regions. Those neighbouring regions which are found to be in the same depth level are considered as a unique region in this segmentation level.

The relative depth of the regions of the grey level segmentation is estimated by considering the occlusions between regions and the motion coherence between neighbouring regions. This estimation procedure can be described in four steps: Motion estimation, motion parameters comparison, overlapping computation and depth level assignation.

4.1. Motion estimation

The motion estimation is based on the regions produced by the grey level segmentation. These regions are assumed to define parts of objects of the scene. Thus, they undergo a coherent motion. A parametric motion model is defined for these regions. In this work, the parameters of the affine motion model are estimated for every region at frame t, using the information of frame t-1. Different techniques can be used to estimate this motion. In this implementation, the method proposed in [2] has been used.

4.2 Motion parameters comparison

The motion parameters between neighbouring regions are compared in order to detect those regions which move with coherent motion. For this aim, the motion fields obtained for the two regions are approximated using the same model parameters for the union of the two regions. The motion similarity measure is taken as the maximum in the two regions of the mean absolute error between the original motion field in the region and the one obtained in this way.

4.3 Overlapping computation

The clues for the depth estimation are obtained from the detection of the covered and uncovered zones. Covered zones correspond to overlapping zones once the regions have been motion compensated. When there is relative motion between two neighbouring regions (A and B) belonging to different depth levels, an overlapping zone appears at frame t. Overlapping zone means that pixels of frame t-1 and region A compensated with its estimated motion achieve the same position at frame t as pixels from region B compensated with its corresponding motion. To decide which one of the two regions is in the foreground it has to be checked whether the pixels of this overlapping zone belong to region A or B at frame t. This process is illustrated in Figure 2.

We can also extract information from the uncovered zones. These zones correspond to an overlapping zone if we invert the order of the images. That is, the uncovered zones becomes overlapping zones if we examine the motion from t to t-1.

In order to decide to which region belongs the overlapping zone, the compensation error in this zone is computed using two different hypothesis:

- The overlapping zone belongs to region A: motion parameters of region A are use to obtain the mean absolute compensation error per pixel.
- The overlapping zone belongs to region B: motion parameters of region B are used to obtain the mean absolute compensation error per pixel.

The compensation error obtained using the first hypothesis is stored in a matrix *comp_err(A,B)* and *comp_err(B,A)* for the second hypothesis.

The overlapping zone is assumed to belong to the region which produces a lower compensation error. Thus, the region to which this zone belongs is temporary assigned to the foreground of the other one. However, this assumption is more reliable in some cases than others. In particular, it will be more reliable if the difference in compensation error with the two hypothesis considered is large, and if the overlapping zone is large. This reliability will be taken into account in the next step, using the information stored in *comp_err*.

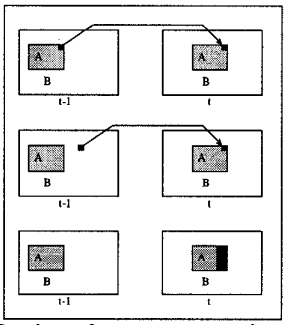

Figure 2. Region A moves to the right in the foreground of B. First and second row: a pixel from region A, compensated with the motion of A achieves the same position than a pixel from region B, compensated with the motion of B. Third row: Overlapping zone.

4.4 Depth level assignation

In the overlapping computation step, an ordering relation between neighbouring regions is obtained if there is a relative motion between them. In this step every region is assigned to a depth level considering this information. For this aim a relaxation labelling algorithm is used [6]. This algorithm is described in the following:

If N different labels for the depth are used (being label 1 foreground and label N background), every region i is assigned an initial probability of being at label λ, $p_i(\lambda)$. These probabilities are initialised to 1/N for the first image and to the final probability achieved at image t-1 for image t. Those final probabilities must be slightly modified in order not to start from too low probabilities at image t (note that, from the following

equation, $p_i^{k+1}(\lambda)=0$ if $p_i^k(\lambda)=0$). These probabilities are iteratively updated using:

$$p_i^{k+1}(\lambda)=\frac{p_i^k(\lambda)\left(1+q_i^k(\lambda)\right)}{\sum_{\lambda} p_i^k(\lambda)\left(1+q_i^k(\lambda)\right)}$$

$$q_i^k(\lambda)=\sum_{j\in Neigh(i)} c_{ij}\left(\sum_{\alpha=1}^{N} r_{ij}(\lambda,\lambda_\alpha)p_j^k(\lambda_\alpha)\right)$$

where k is the iteration number and $r_{ij}(\lambda,\lambda_\alpha)$ is the compatibility between label λ on region i and level λ_α at region j. This compatibility is 1 if region i is in the foreground of j and λ is smaller than λ_α or if it is in the background and λ is larger than λ_α. The compatibility is (-1) in any other case.

Neigh(i) stands for the set of neighbouring regions of region i.

c_{ij} is the relative certainty of the order relation between regions i and j. This information is extracted from the matrix *comp_err* and is given by different factors, as the size of the overlapping zones and the relation between *comp_err(i,j)* and *comp_err(j,i)*.

When the information about the overlapping zone is not reliable enough (c_{ij}<threshold) it is usually due to the fact that the two neighbouring regions are in the same depth level and follow the same motion. Thus, the information extracted in the motion parameters comparison step is used. If the two regions are very similar in motion, then $r_{ij}(\lambda,\lambda_\alpha)$ is set to 1 if $\lambda=\lambda_\alpha$ and (–1) if $\lambda\neq\lambda_\alpha$. c_{ij} will be 1 if the two sets of parameters are equal and it will decrease if the similarity is not so large. In this way we will increase the probability of two neighbouring regions being at the same depth level if they have a similar motion. There is also the possibility that these two regions belong to different depth levels but have similar motion. In this case there is no information to obtain the real depth level, and the error would be corrected as soon as there is relative motion between the objects in the sequence.

We can observe that this relaxation algorithm has the following properties:

- The probability $p_i(\lambda)$ of region i having the label λ is increased if other neighbouring regions have high probability of being at labels compatible with λ being at region i.
- $p_i(\lambda)$ is decreased if other high probability labels are incompatible with λ at i.
- Labels with low probability have little influence on $p_i(\lambda)$, whether or not they are compatible with it.
- If the relative certainty between regions (c_{ij}) is high, the influence of its compatibility on the new probabilities is also high.
- The initialisation of the probabilities to the final probability achieved in the previous image guaranties a temporal stability. Only if there is a highly reliable new information the label will be changed.

Convergence usually occurs after a few iterations, and every region is assigned to the depth level λ which maximises $p_i(\lambda)$. Regions can also be assigned to an uncertainty zone when there is not information enough to assign them to a particular depth level.

5. Results

In Figure 3 and 4 two examples of the segmentation method applied to two sequences is described. In the first row two original images of the sequence are shown. In the second row we have the partition produced for these two images by the bottom level segmentation. Every region is identified by a different grey level value. The last image of this row shows the decision about the depth level of every region. In this example only two levels are found. Regions in black correspond to the foreground, while regions in white correspond to the background. Following this classification step, clusters of regions with the same depth level would be merged to create the top level of the segmentation.

6. Conclusions

The segmentation scheme presented uses a relative depth estimation algorithm in order to introduce higher level information. Using the information about the occlusions and the motion coherence between regions only the relation between pairs of neighbouring regions can be obtained. The application of a relaxation algorithm allows the assignment of every region to a depth level. The higher level information obtained with this procedure is useful to extract the real objects of the scene, which can be used for content-based applications, such as object oriented video coding or object manipulation.

7. References

[1] F. Dufaux, F. Moscheni, "Segmentation-based motion estimation for second generation video coding techniques", Video Coding, The Second Generation Approach, de. by L.

Torres and M. Kunt, Kluwer Academic Publishers, 1996, pp. 219-263.
[2] J.L. Dugelay, H. Sanson, "Differential methods for the identification of 2D and 3D motion models in image sequences", in Signal Processing, Image Communication, 7, 1995, pp. 105-127.
[3] F. Marqués, M. Pardàs, P. Salembier, "Coding-oriented segmentation of video sequences", in Video Coding, The Second Generation Approach, edited by L. Torres and M. Kunt, Kluwer Academic Publishers, 1996, pp. 79-123.
[4] M. Pardàs, P. Salembier,"*Time recursive segmentation of image sequences*", EUSIPCO-94, VII European Signal Processing Conference, Edinbourgh, September 1994, pp. 18-21.
[5] M. Pardàs, "Relative depth estimation and segmentation in monocular schemes", Picture Coding Symposium, PCS 97, Berlín, Germany, September 1997, pp. 367-372.
[6] A. Rosenfeld, R. Hummel, S. Zucker, "Scene labeling by relaxation operations", IEEE Transactions on systems, man and cybernetics, 6, 1976, pp. 420-433.
[7] P. Salembier, F. Marqués, M. Pardàs. "Segmentation-based video coding: temporal links and rate control". Proceedings of EUSIPCO-96, Trieste, Italy, September 1996, pp. 455-458.

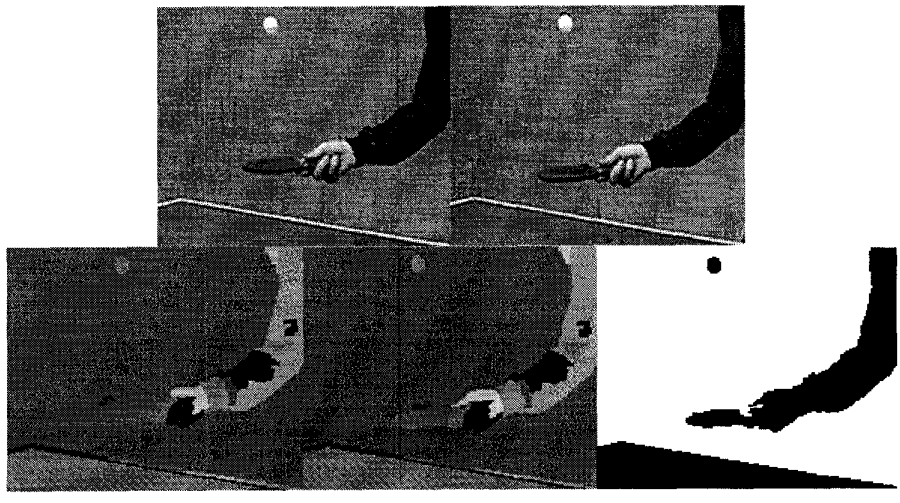

Figure 3. First row: Two original images from the sequence Table Tennis. Second row: Grey level segmentation and depth segmentation.

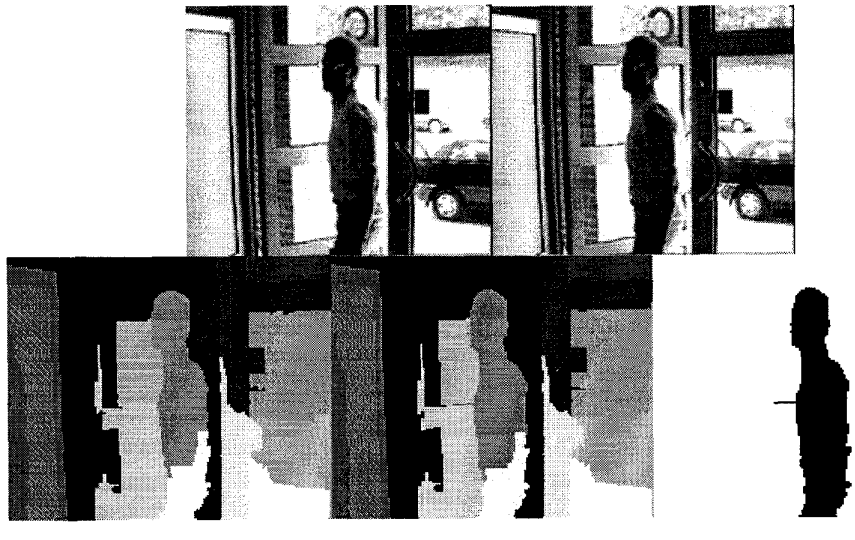

Figure 4. First row: Two original images from the sequence "entrance_inside". Second row: Grey level segmentation and depth segmentation.